Statistics and Data Science 171

YData: Text Data Science

Overview of Topic Models

February 21

Intro to Topic Modeling

Some of the following slides are from Dave Blei's tutorial on Topic Modeling

`http://www.cs.columbia.edu/~blei/topicmodeling.html`

A survey paper describing many of these ideas in more detail is here:

`https://cacm.acm.org/magazines/2012/4/147361-probabilistic-topic-models/fulltext`

Topic modeling

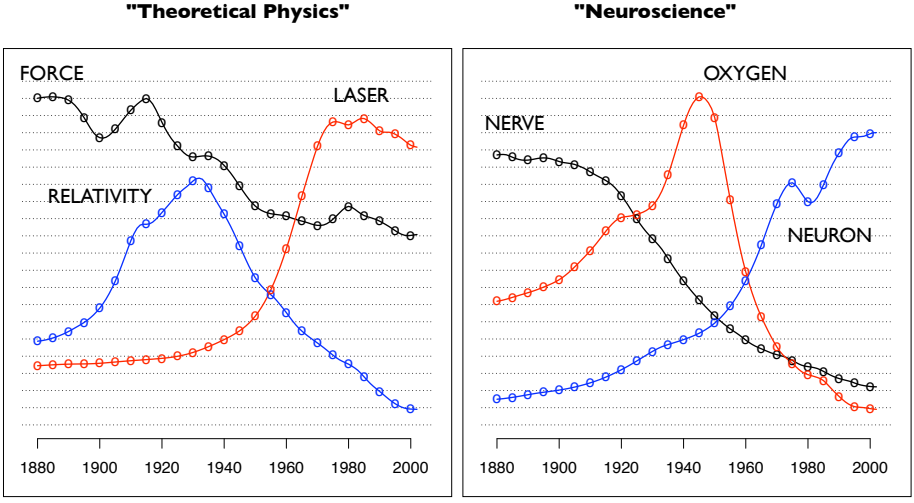
Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

1. Discover the hidden themes that pervade the collection.
2. Annotate the documents according to those themes.
3. Use annotations to organize, summarize, and search the texts.

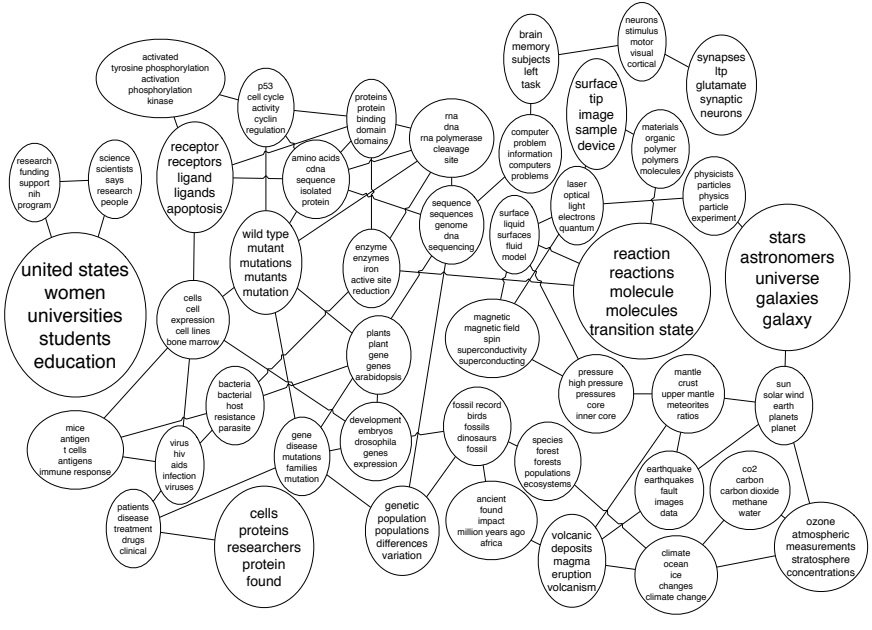
Discover topics from a corpus

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

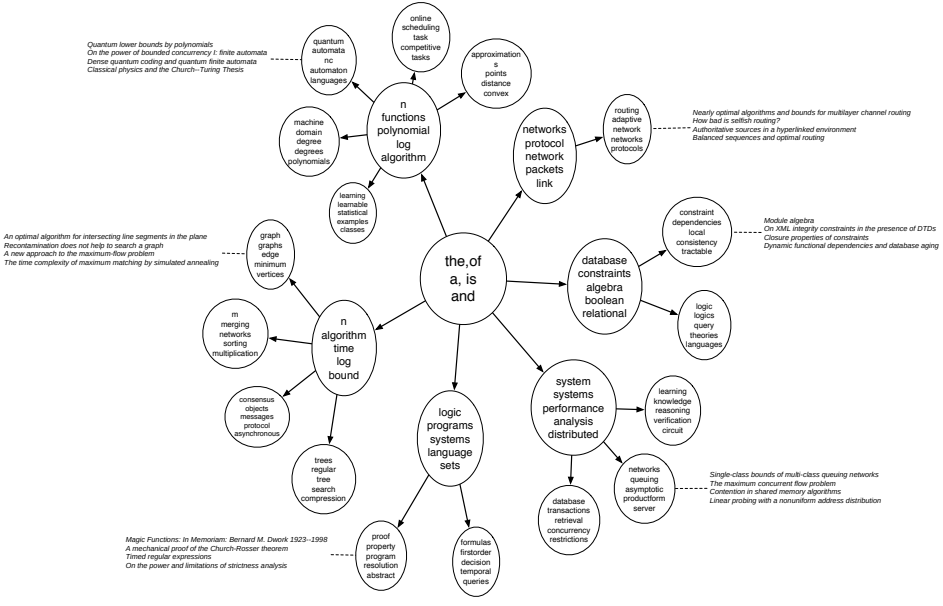
Model the evolution of topics over time



Model connections between topics



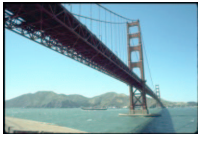
Find hierarchies of topics



Annotate images

SKY WATER TREE
MOUNTAIN PEOPLE

SCOTLAND WATER
FLOWER HILLS TREE



SKY WATER BUILDING
PEOPLE WATER

FISH WATER OCEAN
TREE CORAL

PEOPLE MARKET PATTERN
TEXTILE DISPLAY

BIRDS NEST TREE
BRANCH LEAVES

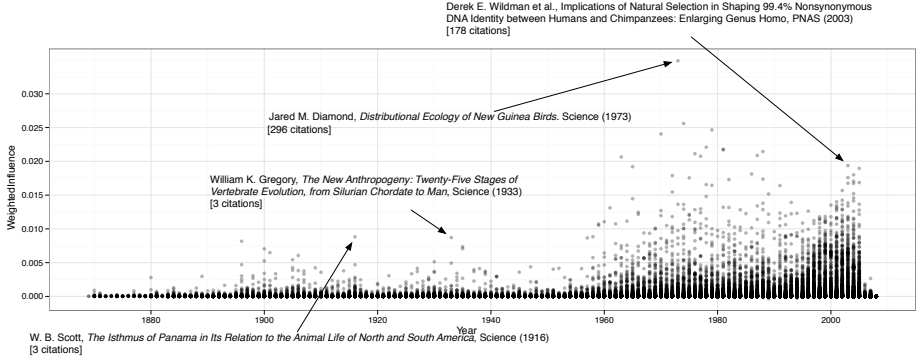
Captioning

there is a large bird on the water
a small bird sitting on top of a lake
a large white bird standing on the water on a beach
a bird is on the water on a beach
a bird that is standing in the water

a professional baseball game is played in the middle of the field
several players at the end of a baseball game
a group of players playing a baseball game
the baseball players are playing games at the field
a baseball players are playing with a game and fans

www.cs.toronto.edu/~nitish/nips2014demo/

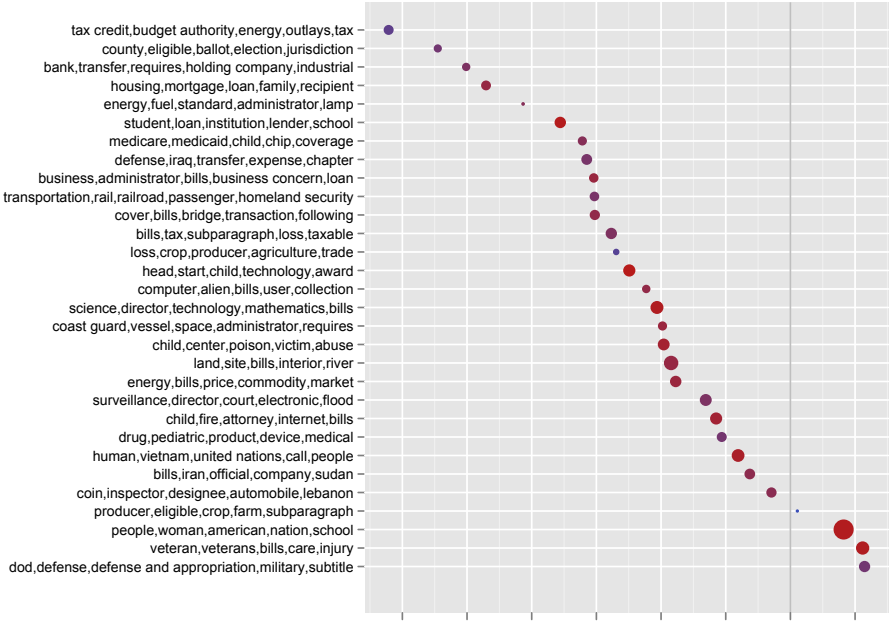
Discover influential articles



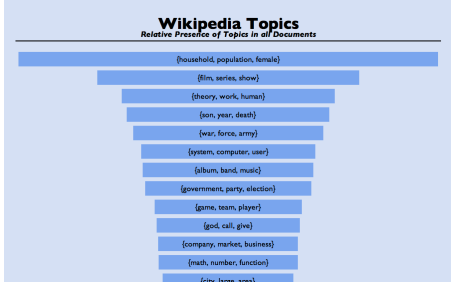
Predict links between articles

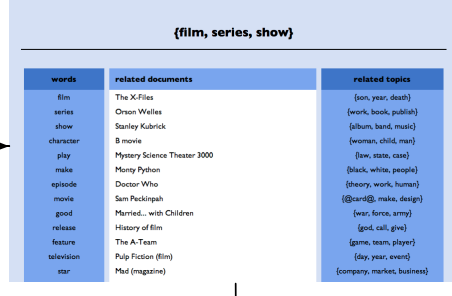
Markov chain Monte Carlo convergence diagnostics: A comparative review	
Minorization conditions and convergence rates for Markov chain Monte Carlo	**RTM (ψ_e)**
Rates of convergence of the Hastings and Metropolis algorithms	
Possible biases induced by MCMC convergence diagnostics	
Bounding convergence time of the Gibbs sampler in Bayesian image restoration	
Self regenerative Markov chain Monte Carlo	
Auxiliary variable methods for Markov chain Monte Carlo with applications	
Rate of Convergence of the Gibbs Sampler by Gaussian Approximation	
Diagnosing convergence of Markov chain Monte Carlo algorithms	
Exact Bound for the Convergence of Metropolis Chains	**LDA + Regression**
Self regenerative Markov chain Monte Carlo	
Minorization conditions and convergence rates for Markov chain Monte Carlo	
Gibbs-markov models	
Auxiliary variable methods for Markov chain Monte Carlo with applications	
Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Models	
Mediating instrumental variables	
A qualitative framework for probabilistic inference	
Adaptation for Self Regenerative MCMC	

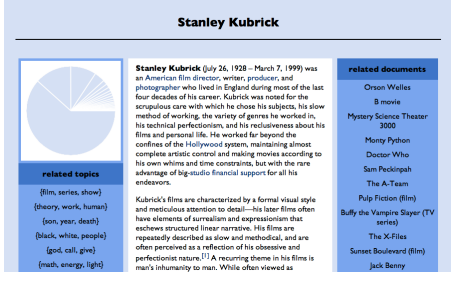
Characterize political decisions

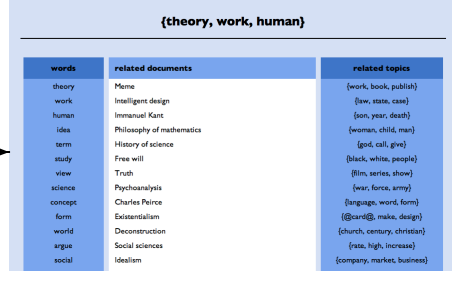


Organize and browse large corpora









Topics in scientific texts

Quantum physics	spin energy field electron magnetic state states hamiltonian
Particle physics	higgs neutrino coupling decay scale masses mixing quark
Astrophysics	mass gas star stellar galaxies disk halo radius luminosity
Relativity	black metric hole schwarzschild gravity holes einstein
Number theory	prime integer numbers conjecture integers degree modulo
Graph theory	graph vertex vertices edges node edge number set tree
Linear algebra	matrix matrices vector basis vectors diagonal rank linear
Optimization	problem optimization algorithm function solution gradient
Probability	random probability distribution process measure time
Machine learning	layer word image feature sentence model cnn lstm training

Topics modeling for equations

Topic	Generated Equations
Quantum physics	• $E = \hbar \frac{\partial^2 S}{\partial t^2}(\frac{\partial \varphi}{\partial c}) - \frac{k}{\hbar^2}\frac{\partial B}{\partial t}\,(t + \partial_t \delta)$. • $\Psi_{\text{pr}} = \sum_{\mathbf{l}}(\psi_{\mathbf{r}+\uparrow} - \psi^{\dagger}_{\mathbf{r}\downarrow}) + \sum_{\mathbf{r}'}(\psi^{\dagger}_{\mathbf{r}\downarrow,\uparrow} - \psi_{\mathbf{r}\downarrow}\sigma^{\dagger})$.
Particle physics	• $\mathcal{H} = \frac{2}{4}(\partial_\mu \phi)^2 + 2m\phi_\nu(\phi) + \frac{1}{2}m^2(\phi)(1-\phi^2)^2$. • $m_{\text{eff}}(M) = 1.4 \cdot 10^{-13}\,\text{GeV}$.
Relativity	• $\mathcal{M} = \frac{2}{2}g^{\mu\nu}(f_{\mu\nu,\mu} - g_{\mu\nu,\nu} + g_{\nu\nu,b}f_{\mu,\nu}) + \frac{1}{2}g^{\mu\nu}$. • $T_{\mu\nu} = \int_0^{\infty} ds_{\mu\nu} ds^2 + a_\mu^2 dr^2 + r^2 d\Omega^2$.

Michihiro Yasunaga:

https://michiyasunaga.github.io/files/papers/TopicEq_aaai19_final.pdf

https://seas.yale.edu/news-events/news/michihiro-yasunaga-wins-cra-outstanding-undergraduate-researcher-award

Uber Topics

From a former student:

> *I just wanted to let you know that I'm currently using Latent Dirichlet Allocation at my current job at Uber!*
>
> *We're using LDA to discover topics in rider feedback – when riders write comments about their driver after the trip. We're trying to find topics such as 'unprofessional driver', 'driver no-show', 'sexual harassment', etc. LDA has worked really well with this – so thank you for covering it in much detail in your course.*

Introduction to Topic Modeling

Probabilistic modeling

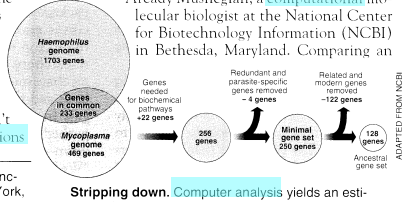
1. Data are assumed to be observed from a generative probabilistic process that includes hidden variables.
   - *In text, the hidden variables are the thematic structure.*
2. Infer the hidden structure using posterior inference
   - *What are the topics that describe this collection?*
3. Situate new data into the estimated model.
   - *How does a new document fit into the topic structure?*

Latent Dirichlet allocation (LDA)

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



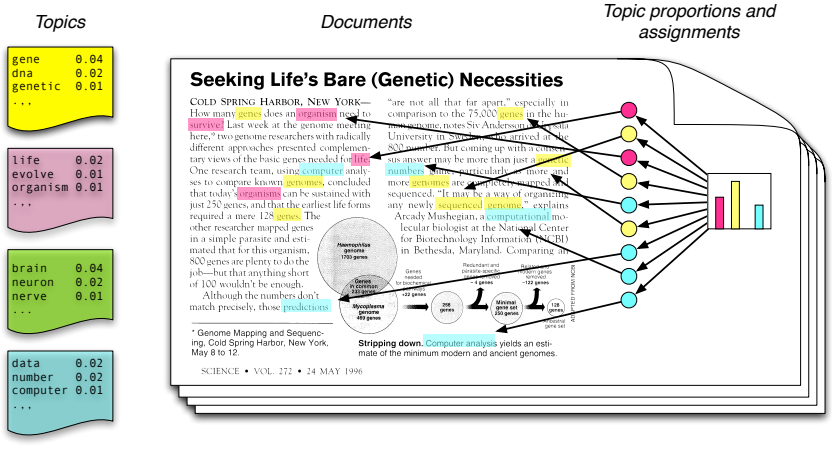
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

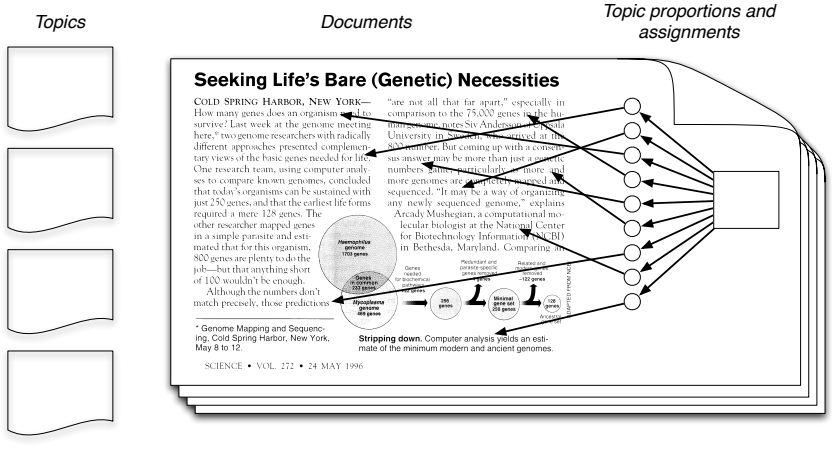
Simple intuition: Documents exhibit multiple topics.

Generative model for LDA



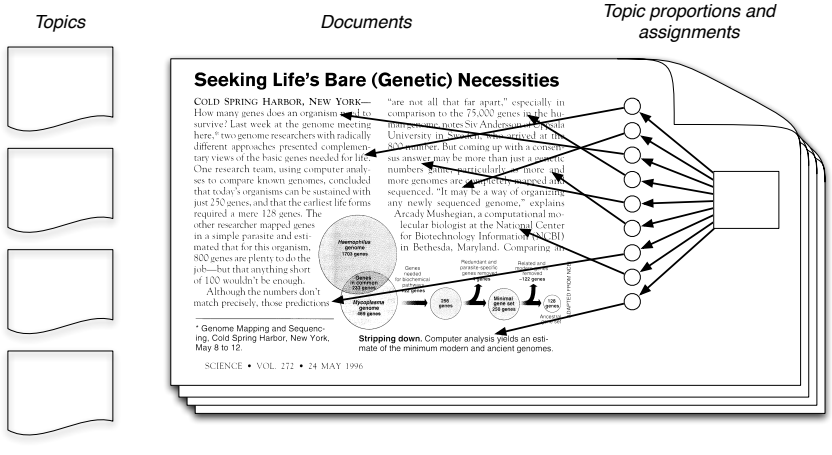
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

The posterior distribution



- In reality, we only observe the documents
- The other structure are **hidden variables**

The posterior distribution



- Our goal is to **infer** the hidden variables
- I.e., compute their distribution conditioned on the documents

$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

Summary

- Topic models automatically extract “semantic themes” from large document collections
- Based on latent variable models
- Can be useful for a wide variety of data